RESOLUTION OF MALDI-TOF COMPARED TO WHOLE GENOME SEQUENCING FOR IDENTIFICATION OF BACILLUS SPECIES ISOLATED FROM CLEANROOMS AT NASA JOHNSON SPACE CENTER

F. Mazhari[1,2], C. L. Castro[4], R. E. Davis[5], M. G. LaMontagne[1], A. B. Regberg[3], [1] Department of Biology and Biotechnology, University of Houston – Clear Lake, Houston, TX, 77058, USA, [2]Jacobs, JETS Contract, NASA Johnson Space Center, 2101 NASA Parkway, Houston TX 77058, [3]Astromaterials Research and Exploration Science Division, NASA Johnson Space Center, 2101 NASA Parkway, Houston TX, 77058, [4]JES Tech, NASA Johnson Space Center Houston, TX 77058, [5]Texas State University, Johnson Space Center 2101 E NASA Pkwy Houston, TX, 77058, USA

Introduction: The Astromaterials Acquisition and Curation Office at NASA Johnson Space Center maintains cleanrooms to archive extraterrestrial materials returned from space exploration missions. Compared to typical built environments, oligotrophic conditions make these facilities inhospitable to microbes [1]. Despite these controls, bacteria and fungi are regularly cultured from these cleanrooms [2]. In particular, *Bacillus* sp. are frequently isolated during routine microbial monitoring [3]. Endospores associated with this genus can survive extreme environments, such as cleanrooms [4]. This microbial contamination may affect the integrity of astromaterials.

In routine monitoring, microbes isolated from these clean rooms are identified using the VITEK2 Compact and 16S rRNA sequencing methods. For closely related *Bacillus* species, these techniques have limited resolving power [5]. Whole genome sequencing (WGS) is the current standard for bacterial identification; however, WGS is costly and time-consuming [6]. Matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) is a low-cost, quick method of identifying microbes and appears capable of differentiating closely related strains of bacterial species [7]. Few studies have compared this proteomics technique to whole genome sequencing.

Methods: Samples were collected from astromaterials cleanrooms according to a microbial monitoring protocol [3]. Fourteen isolates, preliminarily identified as *Bacillus* species, using VITEK2 Compact or 16S sequencing, were selected for further study. Isolates were prepared for MALDI-TOF MS as previously described [3] and spectra were generated at the Proteomics and Mass Spectrometry Core Facility at Pennsylvania State University, (University Park, PA) . Positive-ion mass spectra were acquired on a Bruker Ultraflextreme MALDI TOF/TOF mass spectrometer as previously described [7]. Isolates were prepared for WGS by extracting genomic DNA using the Promega Wizard Genomic DNA Purification Kit (Promega, Madison, WI) according to manufacturer's protocol with an additional lysozyme step. DNA was sequenced

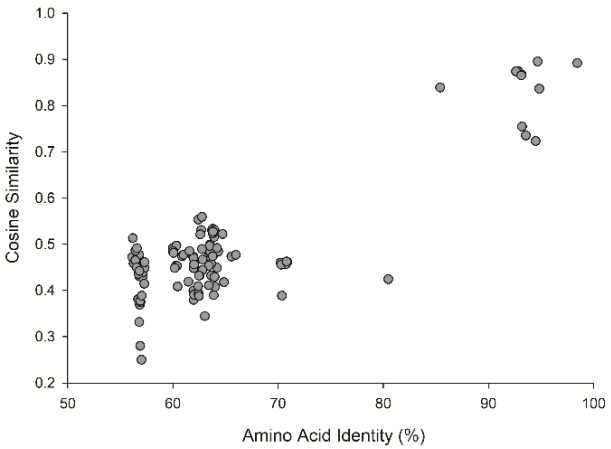

Figure 1. Pairwise similarity of amino acids from single core genes (y-axis) with cosine similarity of mass spectra generated with MALDI-TOF MS.

with a MinION Mk1C from Oxford Nanopore Technologies (Oxford, UK) using Rapid Barcoding Sequencing (SQK-RBK004) according to manufacturer's protocol in two sequence runs consisting of 8 and 12 DNA samples (100-400 ng) using a R.9.4.1 flow cell. Additional short-read sequencing was performed with an Illumina MiSeq using MiSeq Reagent Kit v3 (San Diego, CA) with paired end reads. Raw sequence data was assembled on the EDGE Bioinformatics v 2.4.1 hybrid genome assembly pipeline [15] using Unicycler [16] for de novo assembly. The closest bacterial strain identity was determined using the Type (Strain) Genome Server (TYGS) [8] by comparing all genomes with available strain genomes in the database and by extracting the 16S rRNA sequence and aligning against the database. Pairwise similarity of amino acids was generated with EzAAI with default parameters [9]. A correlation graph was generated in Rstudio [14] using pairwise similarity scores for mass spectra generated by MALDI-TOF MS as described previously [7] and pairwise similarity of amino acid sequences was predicted from single core genes in draft genomes.

Results and Discussion: In this study, draft genomes were generated from 14 strains of *Bacillus*. These isolates exhibited genomic similarities to strains frequently found in other cleanroom facilities used for spacecraft assembly. When compared to the 16S rRNA sequencing, phylogenomic analysis of single-copy core

genes showed a higher taxonomic resolution and revealed more species diversity. The TYGZ type strain database indicates that 5 isolates are potential new species. The VITEK2 Compact was unable to identify 3 isolates at the genus level and all 14 isolates at the species level.

Mass spectra from MALDI-TOF MS showed good agreement with WGS (Figure 1). Pairs of strains that were > 94% similar to each other in terms of predicted amino acid sequences consistently were > 0.65 similar to each other in terms of mass spectra. This degree of similarity is consistent with the definition of MALDI-TOF taxonomic unit (MTUs) [7]. This suggests that MALDI-TOF and WGS are consistent with one another and that MALDI-TOF can quickly, and with a resolution comparable to WGS, identify strains of *Bacillus* sp. isolated from cleanroom environments. These findings also point to the existence of a cosmopolitan class of *Bacillus* species that are frequently found in cleanrooms and similar built environments.

It is important to identify organisms at the strain-level in low biomass environments because different strains may have divergent metabolic strategies that infer specialized survival mechanisms. Current techniques fail to account for species-level diversity and novel strains. Compared to conventional sequencing and biochemical identification methods, MALDI-TOF MS is quick, precise, and economical. This, along with its capacity for strain-level resolution, makes it ideal for routine microbial monitoring. However, mass spectral databases used by MALDI-TOF MS systems contain a poor representation of environmental bacteria. Implementation of this method will require building a database of mass spectra from representative MTUs.

References: [1.] McCubbin et al. (2019) Sp. Sci Rev doi:10.1007/s11214-019-0615-9. [2] Regberg et al. (2018) 42nd COSPAR Scientific Assembly. Abstract id. PPP.3-18-18. [3] Mazhari, Farnaz. 2021. Master's Thesis, University of Houston - Clear Lake. [4] Schuerger et al. (2003) *Icarus* 165:253-76. [5] Seuylemezian *Frontiers in Microbiology* 9:e780. [6] Didelot et al. (2012) *Nature Reviews Genetics* 13:601-612. [7] LaMontagne et al. (2021) PeerJ 9:e11359. [8] Meier-Kolthoff et al. (2019) *Nat Commun*, 10(1):2182. [9] Kim et al. (2021) *J Microbiol*. 59,476–480. [10] Kotera et al. (2015) Encyclopedia of Metagenomics. https://doi.org/10.1007/978-1-4899-7478-5_694 [11] Stamatakis et al. (2014) *Bioinformatics*. 30(9):1312-3. [12] Lee et al. (2019) *Bioinformatics*, Volume 35, Issue 20, Pages 4162–4164. [13] Letunic et al. (2021) *Nucleic Acids Research*, Volume 49, Issue W1, Pages W293–W296. [14] RStudio Team (2020). PBC, Boston, MA URL http://www.rstudio.com/. [15] Li et al. (2016) Nucleic Acids Research. [16] Wick et al. (2017) PLoS Comput Biol.